The relationship of Yellow card and Full-Time Away Team Goals*

Analysis of Primer league 2023-24 season

Che-Yu Wang

April 17, 2024

Football match outcomes, especially Full-Time Away Team Goals (FTAG), are difficult to predict in sports analytics. This study uses 2024 Premier League data to examine how Home Team Yellow Cards (HY) affect away team scoring. The study uses logistic, Poisson, and negative binomial regressions to find the best statistical model for football score count data. Poisson regression, ideal for count data, assumes equal mean and variance, which is often violated in real-world data. Logistic regression models binary outcomes and may overlook nuanced count data. Negative binomial regression better represents data variability by adding an overdispersion parameter. The study evaluates each model's ability to capture football scoring patterns' complexities using statistical analysis, model comparison, and diagnostic measures. The findings emphasize the delicate balance between model complexity and predictive accuracy, emphasizing the importance of choosing the right model to understand away team goal-scoring trends. This meticulous approach improves football analytics predictive modelling and sports statistics by refining count-based outcome analysis..

1 Introduction

By 1900, a significant portion of the male population spent leisure time playing or watching sports, with football being a popular choice due to its spontaneity and order. The Victorian period marked the separation between work and leisure, with industrial work rhythms and rapid urban growth causing a shift in behavior and promoting leisure (Mason 2023). The Premier League is the highest level of the English football league system. Contested by 20 clubs, it operates on a system of promotion and relegation with the English Football League.

*Code and data are available at:https://github.com/Cheyuwang/Primer-league-analysis.

The prediction of football match outcomes, particularly Full-Time Away Team Goals (FTAG), is a major challenge in the field of sports analytics. This challenge is driven not only by the inherent unpredictability of football matches, but also by a growing interest in understanding and forecasting sports events among academics and industry professionals. The ability to accurately predict these outcomes has important implications for strategic planning in sports management, the betting industry, and fan engagement. Despite the abundance of research in this area, football's dynamic and complex nature, characterized by unpredictable gameplay and a plethora of influencing factors, presents a persistent challenge in achieving high prediction accuracy. This gap emphasizes the importance of continuously refining predictive models capable of handling the complexities of football data.

This research uses a large-scale dataset from the football-data.co.uk website (soccer-Data 2024) and focuses on the 2024 Premier League season. We use R Core Team (2023), UK (2024) to completed and hlep our analysis. Several metrics are included in the dataset, such as Home Team Yellow Cards (HY), which are thought to affect the away team's capacity for scoring. In the context of football match outcomes, the study attempts to analyze and comprehend the predictive power of logistic, Poisson, and negative binomial regression models. Every model offers a unique method for managing count data: logistic regression models binary outcomes, Poisson regression handles count data with equal variance and mean, and negative binomial regression adds a parameter to handle overdispersion.

The comparative analysis of the study shows that, although logistic regression is a straightforward technique for forecasting discrete events, it might not make full use of the count data that is available, which could lead to information underutilization. Although poisson regression works well with count data, it often breaks down because it assumes equal mean and variance, which is rarely the case with football match results. Among the models that can better explain the variability and over dispersion found in football scoring data is negative binomial regression. The results indicate that negative binomial regression has a great deal of potential for increasing the accuracy of football match outcome predictions because they highlight the delicate balance between model complexity and predictive performance.

This study's contribution to the developing field of sports analytics—which offers knowledge that can greatly enhance football match prediction modeling—makes it noteworthy. The study is a useful tool for analysts, researchers, and practitioners of sports analytics because it identifies the best statistical model for FTAG prediction. The study also lays the foundation for future research by highlighting the significance of model selection in comprehending and predicting athletic events. The structure of the paper is as follows: After the introduction, Section 2 reviews the literature and provides an overview of the field's prior research as well as its theoretical foundations. The methodology—which covers the dataset, variables taken into account, and statistical models employed—is covered in Section 3. The results and analysis, which contrast the predictive powers of logistic, Poisson, and negative binomial regressions, are presented in Section 4. The implications of the results, the limitations of the study, and potential future research directions are covered in Section 5's conclusion.

2 Data

One of the biggest football betting websites in the UK, football-Data.co.uk, is where the data was gathered. (soccer-Data 2024) UK (2024). The website has all of the global football league data, including all of the results from every year with every possible outcome. I analysis the data in R(R Core Team (2023)), with additional tools for support the analysis, including tidyverse(Wickham et al. (2019)), here(Müller (2020)), dplyr(Wickham et al. (2023)), readr(citereadr), modelsummary(Wickham et al. (2023)), janitor(Firke (2023)), tibble(Müller and Wickham (2023)), ggplot2(Wickham (2016)), rstanarm (Goodrich et al. (2022)) and research method is from Alexander (Alexander (2023)),and the article of (Mason and Porter (2023)), (Soares and Shamir (2016)).

2.1 Variables

The primary factors influencing game dynamics in the analysis of football matches are scoring outcomes and disciplinary measures. One important predictor is the variable 'HY', which stands for the number of yellow cards given to the home team. It is hypothesized that yellow cards, a form of discipline for fouls and unsportsmanlike conduct, will affect the performance of the home team and the atmosphere of the match. The dependent variable 'FTAG' represents the number of goals scored by the opposing team at the conclusion of the game. The study investigates the connection between the away team's scoring success and the home team's disciplinary actions using a Poisson regression model. This analysis can guide coaching and game preparation tactics and is essential for comprehending team behaviour.

2.2 Data cleaning preparation

The analysis relied on a dataset from a reputable football statistics website, ensuring high accuracy and completeness. The variables 'HY' (Home Yellow Cards) and 'FTAG' (Full Time Away Goals) were highlighted as critical for understanding disciplinary actions and their impact on match outcomes. These variables have a count data structure, making them ideal for Poisson regression. This statistical method accurately models the probability of a number of events occurring in a given period of time, assuming a known constant mean rate that is independent of time since the last event. The Poisson distribution was chosen because it accurately models the probability of events occurring over a fixed period.

2.3 Visualizations and Summary Statistics

In this part, I choose a better understanding in comprehension of the dataset. The histogram in Figure 1 shows the frequency of football matches categorized by the number of yellow cards awarded to home teams. The graph shows a clear mode at 1 yellow card, indicating that most

matches involve a single yellow card. The frequency decreases as the number of yellow cards increases, with two yellow cards being the second most common occurrence. A progressive decline for three or more cards suggests a general discipline level in teams when playing at home, with a majority avoiding high counts of yellow cards. Matches with no yellow cards are less frequent than those with one or two, suggesting at least one caution per game. The instances of 4, 5, and 6 yellow cards are rarer, highlighting outliers in typical in-game conduct. The sharp decrease after three yellow cards suggests that it is uncommon for home teams to receive a high number of cautions in a single match, potentially indicating unusually aggressive or unruly play. The table 1 displays data on football match analysis, such as disciplinary actions ('HY' for Home Yellow Cards) and team identifiers ('FTAG'). However, instead of team names, the 'FTAG' column is expected to contain Full Time Away Goals, a numerical variable. This inconsistency indicates either a mislabeling or incorrect data representation. In an optimal state, 'HY' would quantify the number of yellow cards issued to the home team during individual matches, whereas 'FTAG' would quantify the away team's success rate in scoring goals at full-time. This would allow for a thorough investigation of the relationship between home team discipline and the opposing team's scoring ability. A verification and correction process is required to align 'FTAG' with the numerical goal data. Once accurate, the dataset will be fed into a Poisson regression model, which could reveal dynamics in football matches and potentially influence game strategy and team behaviour regulation.

2.4 Alternative Data Considerations

In this part, I choose a better understanding in comprehension of the dataset. The histogram in Figure 1 shows the frequency of football matches categorized by the number of yellow cards awarded to home teams. The graph shows a clear mode at 1 yellow card, indicating that most matches involve a single yellow card. The frequency decreases as the number of yellow cards increases, with two yellow cards being the second most common occurrence. A progressive decline for three or more cards suggests a general discipline level in teams when playing at home, with a majority avoiding high counts of yellow cards. Matches with no yellow cards are less frequent than those with one or two, suggesting at least one caution per game. The instances of 4, 5, and 6 yellow cards are rarer, highlighting outliers in typical in-game conduct. The sharp decrease after three yellow cards suggests that it is uncommon for home teams to receive a high number of cautions in a single match, potentially indicating unusually aggressive or unruly play. The Table 1 displays data on football match analysis, such as disciplinary actions ('HY' for Home Yellow Cards) and team identifiers ('FTAG'). However, instead of team names, the 'FTAG' column is expected to contain Full Time Away Goals, a numerical variable. This inconsistency indicates either a mislabeling or incorrect data representation. In an optimal state, 'HY' would quantify the number of yellow cards issued to the home team during individual matches, whereas 'FTAG' would quantify the away team's success rate in scoring goals at full-time. This would allow for a thorough investigation of the relationship between home team discipline and the opposing team's scoring ability. A verification and correction process is required to align 'FTAG' with the numerical goal data. Once accurate,

the dataset will be fed into a Poisson regression model, which could reveal dynamics in football matches and potentially influence game strategy and team behaviour regulation.

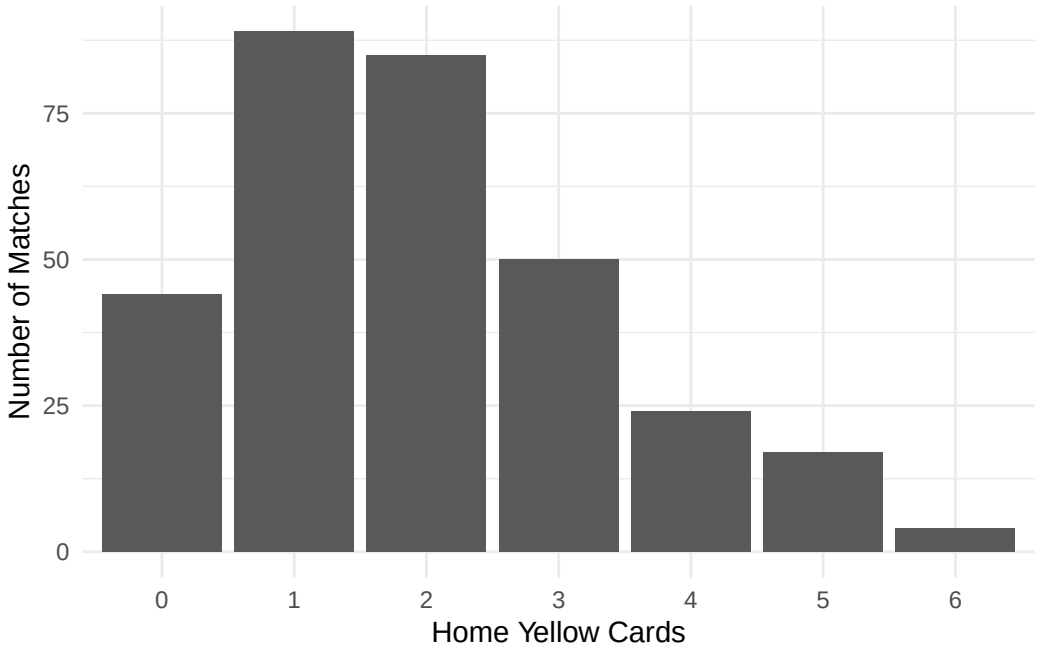

Figure 1: Figure 1

Table 1: Yellow card with Primer league team in away match

```
# A tibble: 19 x 2
      HY premier_league_teams
   <dbl> <chr>
 1     0 Arsenal
 2     1 Aston Villa
 3     2 Brentford
 4     3 Brighton
 5     4 Chelsea
 6     5 Crystal Palace
 7     0 Everton
 8     1 Fulham
 9     2 Leeds United
10     3 Leicester City
11     4 Liverpool
12     5 Manchester City
13     0 Manchester United
14     1 Newcastle United
15     2 Nottingham Forest
16     3 Southampton
17     4 Tottenham Hotspur
18     5 West Ham United
19     0 Wolverhampton Wanderers
```

3 Measurement

Two variables make up the dataset: 'HY', which stands for Home Yellow Cards, and 'FTAG', which records visiting teams' scoring prowess. "HY" measures the quantity of warnings for infractions or unsportsmanlike behaviour that referees have given to the home team; these warnings are documented in official logs and validated by the relevant football associations. These incidents are painstakingly transformed into data points so that the disciplinary actions taken during several matches can be combined and statistically analyzed. Every goal an away team scores is recorded as a unique event in 'FTAG,' which documents the scoring prowess of visiting teams. These goals are frequently the consequence of offensive skill and strategic play. To guarantee a high fidelity representation of each team's scoring record in an away context, these tallies are combined and sent to data curators, such as football statistics websites. As a numerical value that corresponds to the 'FTAG' column in the dataset, the raw goal count represents the outcome of a sequence of events and reactions that take place in the turbulent atmosphere of a football game. Both variables require rigorous recording, validation, and data entry protocols in order to transform transient and complex real-world events into structured data that is ready for analysis.

4 Model

Poisson analysis is a statistical method for modelling count data, with the outcome variable representing the number of occurrences of a specific event within a given unit of time, area, or volume. This approach assumes that events occur independently at a constant average rate throughout the duration or space of interest. The Poisson distribution, with the parameter lambda (λ), describes the probability of observing a certain number of events within a given interval. It is versatile and useful in a variety of fields, including epidemiology, finance, telecommunications, and sports analytics. By applying Poisson models to count data, analysts can uncover underlying patterns, identify factors that influence event frequency, and make informed predictions about future occurrences.

4.1 Model set-up

$ _1 Normal(0, 2.5) $ These priors are normally distributed with a mean of 0, indicating no initial bias towards either an increase or decrease in the count of `FTAG`. - A standard deviation of 2.5 for these priors reflects a moderate level of uncertainty about the coefficients. The probability mass function of the Poisson distribution is as follows, giving the probability of observing exactly events:

$$P_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad \text{for } k = 0, 1, 2, ...$$

Here, (k) denotes the probability of observing exactly k events (in our case, the number of goals scored, denoted as FTAG) in a given time frame. The parameter $ { }$ represents the average rate at which these events occur, and it also serves as the variance of the distribution

- $P_\lambda(k)$: The probability of observing k `FTAG`.
- e: The base of the natural logarithm.
- $\frac{\lambda^k}{k!}$: The probability mass function of the Poisson distribution, where $k!$ denotes the factorial of k

The factorial function, denoted by $k!$, is the product of all positive integers up to k. This function in the denominator normalizes the distribution to ensure that the sum of probabilities across all possible counts of `FTAG` equals 1. When interpreting the results of this model, if β_1 is significantly different from zero, it suggests that there is a statistically significant association between `HY` and the expected count of `FTAG`. A positive β_1 indicates that as `HY` increases, the expected count of `FTAG` increases, and vice versa.

$$
\begin{aligned}
y_i|\lambda_i &\sim \text{Poisson}(\lambda_i) && && (1)\\
\log(\lambda_i) &= \beta_0 + \beta_1 \times \text{Number of words}_i && && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) y_i|\lambda_i && \sim \text{Poisson}(\lambda_i) && (4)\\
\log(\lambda_i) &= \beta_0 + \beta_1 \times \text{department}_i && && (5)\\
P_\lambda(k) &= e^{-\lambda}\frac{\lambda^k}{k!}, \quad \text{for } k = 0, 1, 2, ... && && (6)
\end{aligned}
$$

4.1.1 Model justification

Poisson analysis is a statistical method for modelling count data, with the outcome variable representing the number of occurrences of a specific event within a given unit of time, area, or volume. This approach. Assumes that events occur independently at a constant average rate throughout the duration or space of interest. The Poisson distribution, with the parameter lambda, describes the probability of observing a certain number of events within a given interval. It is versatile and useful in a variety of fields, including epidemiology, finance, telecommunications, and sports analytics. By applying Poisson models to count data, analysts can uncover underlying patterns, identify factors that influence event frequency, and make informed predictions about future occurrences.

5 Results

The Table 2 Poisson regression analysis shows a statistical relationship between the number of yellow cards issued to the home team (HY) and the number of full-time away goals (FTAG) in a set of football matches. The model's intercept represents a baseline rate of away goals when no yellow cards are issued to the home team, implying that away teams have a measurable and significant propensity to score goals regardless of the home team's discipline. However, the coefficient for HY is approximately 0.0421, with a standard error of 0.0323, which exceeds the conventional threshold for statistical significance. This suggests that disciplinary actions measured by yellow cards issued to the home team have little direct impact on the away team's scoring results. It also raises the possibility that factors other than those captured in this simple model may have a greater influence on the number of goals scored by the away team, such as defensive and offensive capabilities, coaching strategic decisions, or players' psychological states. This analysis highlights the complexities of football match dynamics and warns against attributing scoring patterns to single variables like disciplinary actions.These statistics are critical for determining the model's predictive ability and guiding the selection of the best complex model for the data.

The model's fit is moderate, and the influence of home yellow cards on away goals is not pronounced. This suggests that other factors may be more predictive of the number of goals the away team scores, or that the relationship between home yellow cards and away goals is complex and possibly non-linear, which a simple Poisson regression model cannot capture. This lack of a significant relationship suggests the multifactorial nature of football scoring dynamics and the need for a broader investigation into match events and outcomes.

Table 2: Kabble 1 for poisson regresion

term	estimate	std.error	statistic	p.value
(Intercept)	0.3005427	0.0809249	3.713845	0.0002041
HY	0.0421462	0.0322993	1.304864	0.1919391

The bar chart Figure 2 , the Poisson distribution that is used to model the frequency of occurrence in football games is represented visually in a bar chart. It quantifies the difference in away teams' offensive success by classifying matches according to full-time away goals, which vary from 0 to 7.5. The y-axis measures the frequency of each away goal tally in the dataset by quantifying the matches. The graph also displays a standard Poisson distribution, in which matches with the highest frequency of lower away goals have a progressive decline in goal frequency as the number of goals rises. The chart also shows variation in the amount of yellow cards awarded for various goal totals. Higher yellow card counts do not, however, appear to be correlated with any particular range of away goals, indicating that the dataset does not have a strong linear relationship. To conduct a more thorough analysis, it would be necessary to investigate the significance of the visual patterns that were noticed, like the frequency with which matches ended with one or no yellow cards. One would need to delve deeper into statistical measures like regression analyses, chi-squared goodness of fit tests, and variance to mean ratio in order to fully comprehend the dynamics at play. This figure offers a concise, preliminary representation of the distribution of the data among the variables under investigation.

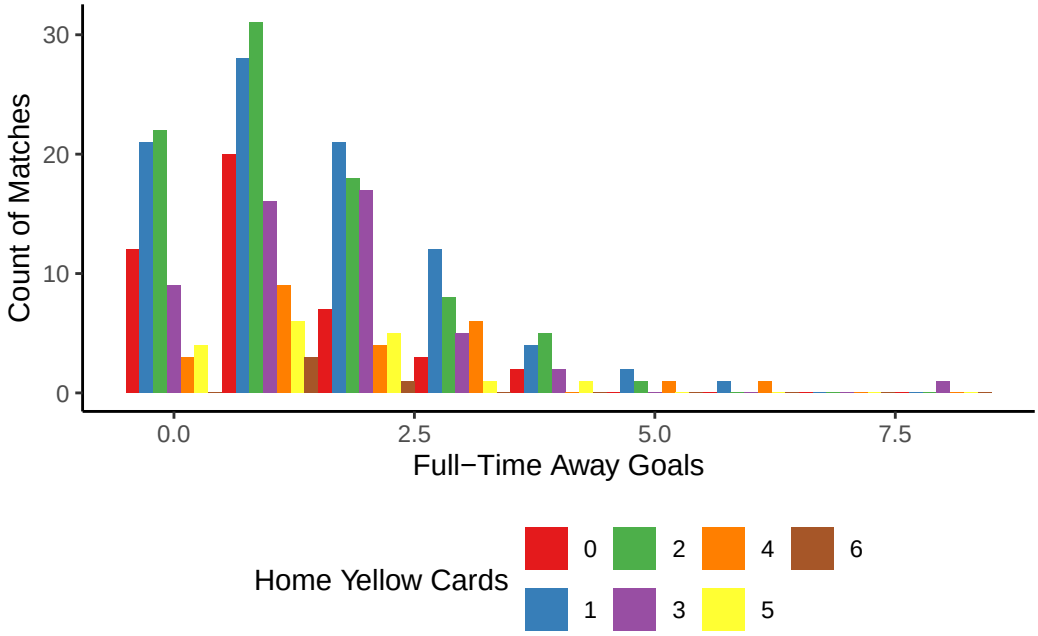

Figure 2: Bar Chart for poisson regresion

6 Discussion

6.1 Finding

This paper investigates the relationship between disciplinary actions, such as yellow cards issued to home teams, and away teams' subsequent scoring success during full-time matches in association football. The study uses a dataset from football-data.co.uk, a comprehensive source of match statistics, and Poisson regression analysis to investigate the impact of home team yellow cards on away team goals. The hypothesis is that increased disciplinary action against the home team may result in more scoring opportunities for the opposing team. A visual analytical approach is also used, with a Poisson distribution bar chart used to visually identify patterns or anomalies in the distribution of match results. The paper aims to bridge the gap between statistical analysis and practical insights into football match outcomes by combining rigorous statistical modelling with user-friendly visualization techniques. This empirical study is methodologically sound and easily interpretable by a wide range of readers, including those interested in sports analytics, team strategy, and predictive modelling in sports contexts. The findings aim to foster a better understanding of how on-field events such as yellow cards can influence the flow and eventual outcome of football games, adding to the ongoing discussion about quantitative analysis of sports data.

6.2 Other outcome : Referees decision on yellow card

The study by Jimmy Tanamati Soares and Lior Shamir examines the impact of team reputation factors, such as team rank, budget, and home game audience size, on referee decisions regarding penalty kicks and yellow cards in soccer (Soares, J. T., & Shamir, L. 2016). The researchers analyzed data from four major European soccer leagues over five seasons, focusing on subjective and potentially biased decisions. They found a significant correlation between the chance of a foul resulting in a yellow card and both the team's rank and budget in leagues like the Bundesliga, suggesting a possible bias in favor of more reputable teams. However, these correlations were not uniformly significant across all leagues. The study also found that the likelihood of a play inside the penalty box leading to a penalty kick was not dependent on the team's budget and rank. The findings support some soccer fans' beliefs that referees may favor certain teams, but they also highlight that most referee decisions are not correlated with team reputation (Soares, J. T., & Shamir, L. 2016). The paper discusses a study that is consistent with Soares and Shamir's research on referee decisions in home games. The study investigates the effect of yellow cards on match outcomes, with a focus on the link between home yellow cards and away goals. Although the study found no statistically significant relationship between home yellow cards and away goals, it does raise questions about the impact of external factors such as team reputation on referee decisions. The paper also investigates how in-game disciplinary measures against home teams, such as yellow cards, relate to away team scoring performance, adding to the discussion about refereeing's impact on sport dynamics. Both studies emphasize the difficulty of isolating refereeing's effects on match results. The paper

calls for more research with a broader set of variables and advanced statistical models to gain a better understanding of the interactions between disciplinary actions, referee bias, team strategy, and their collective impact on the game.

6.3 Implications

Analyzing full-time away goals and yellow cards provides valuable insights into match dynamics, allowing teams to create effective game plans and optimize tactical strategies. This analysis also helps with player management by guiding training plans and development programs. It also helps players improve their discipline and decision-making on the field. Understanding how these factors influence match outcomes improves the narrative and drama of football matches for fans, allowing broadcasters and media outlets to highlight these relationships during match coverage. Fantasy football and betting fans can use this information to inform their predictions and betting strategies, enhancing the viewing experience.

6.4 Limitation

The analysis of full-time away goals and yellow cards in football is limited by its reliance on incomplete or error-prone data, which can introduce bias and affect results reliability. It also overlooks contextual factors like team strategy, match importance, player fatigue, and referee tendencies that could influence the relationship. The analysis also focuses on quantitative aspects of match outcomes, but may not fully capture qualitative aspects like team performance or match intensity. Additionally, its retrospective nature, relying on historical data, does not guarantee predictive accuracy, and real-time factors may influence match outcomes in unpredictable ways.

6.5 Weaknesses and next steps

The study of the relationship between full-time away goals and yellow cards in football games revealed advantages such as the use of Poisson regression for statistical analysis. However, weaknesses include data quality limitations, the need for more comprehensive variable selection and interpretation, model assumptions that are adequate, and transparency in code documentation. Addressing these flaws and seizing improvement opportunities will boost the analysis's credibility and provide useful insights into football match dynamics.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Boca Raton: CRC Press. https://tellingstorieswithdata.com/.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Mason, Tony, and Dilwyn Porter. 2023. *Association Football and English Society, 1863-1915 (Revised Edition)*. Routledge.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://tibble.tidyverse.org/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Soares, Jimmy Tanamati, and Lior Shamir. 2016. "Quantitative Analysis of Penalty Kicks and Yellow Card Referee Decisions in Soccer." *American Journal of Sports Science* 4 (5): 84.

UK, Football Data. 2024. *Football Data UK: Historical Football Results and Betting Odds Data*. https://www.football-data.co.uk.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.